Assessing the Performance of Space Weather Models Using Metrics

K. A. Keller, M. Hesse, L. Rastaetter, M. M. Kuznetsova, A. Falasca

NASA Goddard Space Flight Center

Abstract

- Metrics are one tool for assessing the progress of scientific models in space weather predictions. A scientific metric as defined by the National Space Weather Program has three elements: 1) An output parameter from the model such as density, 2) A satellite or ground-based measurement that can be used for comparison, and 3) A quantifiable parameter that can measure the difference between the model parameter and the measurement. We will present results for heliospheric, inner magnetospheric and ionospheric models. For the heliospheric model, we will compare results to ACE plasma data. For the inner magnetosphere, we will compare the results of the models to LANL geosynchronous satellite data. In the ionosphere, we will compare the computed magnetic perturbations of the ionospheric models to measured values in the Greenland magnetometer chain.

Need for Metrics

- Create objective measure of current capabilities both for scientific and operational needs.
- Measure the improvement of model capabilities over time.
- Provide an objective comparison between models with comparable output.

  Metrics which lead to scores near unity now are useless!

Elements of a Metric

- An output parameter from a model.
  - Density or velocity at a satellite position
- A satellite or ground-based measurement that can be used for comparison.
  - Plasma data from ACE
- A quantifiable norm that assesses the difference between the parameter from the model and the measurement.

Current Metrics

- Density and velocity using ACE.
- Ground magnetic perturbations using data from ground magnetometer chains.
- Particle fluxes at geosynchronous orbits using Los Alamos National Laboratory satellite data.

Heliosphere Metric

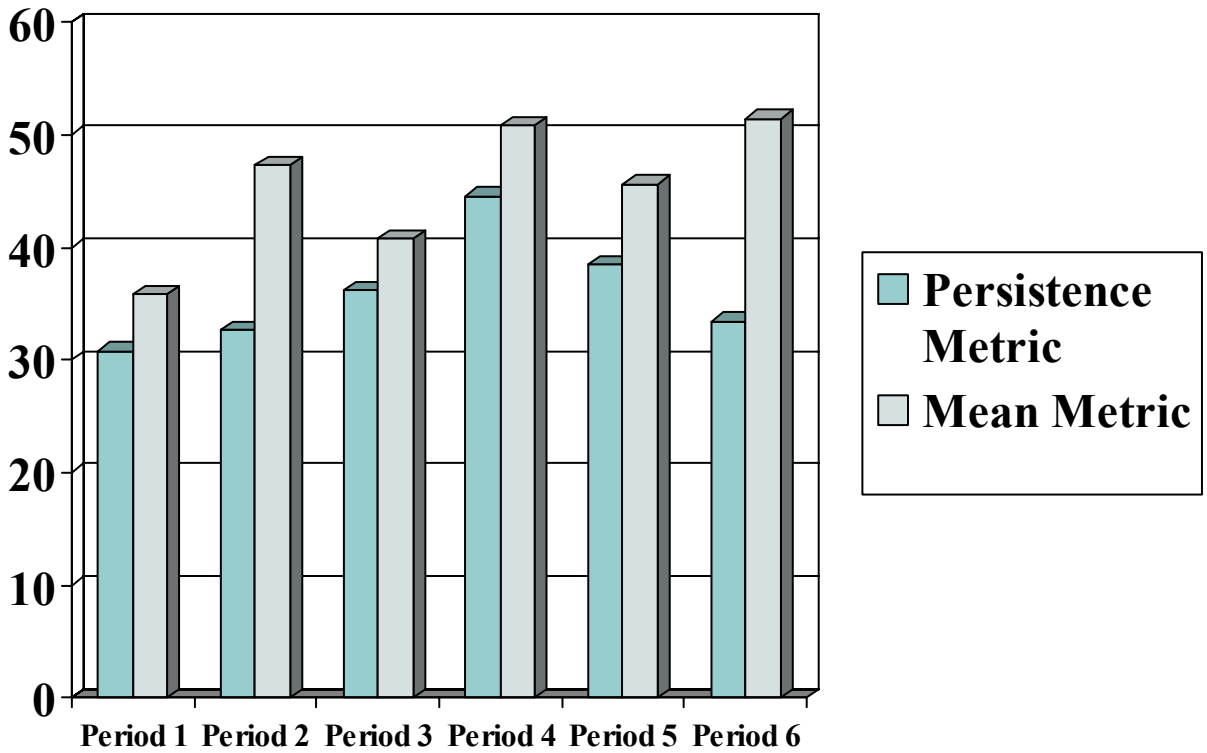
- Data
  - ACE velocity and density average every 6 hours for 27 days.
- Model
  - Heliospheric Tomography Model developed by B. Jackson and P. Hick. This model gives output every 6 hours for 27 days.
- Metric
  - A model is scored using $D_i = \text{sqrt}\,(\Sigma|\Delta H_{model} - \Delta H_{data}|^2/\text{npts})$.
  - A skill score is computed by

$$M_i = 1 - D_i / D_s$$

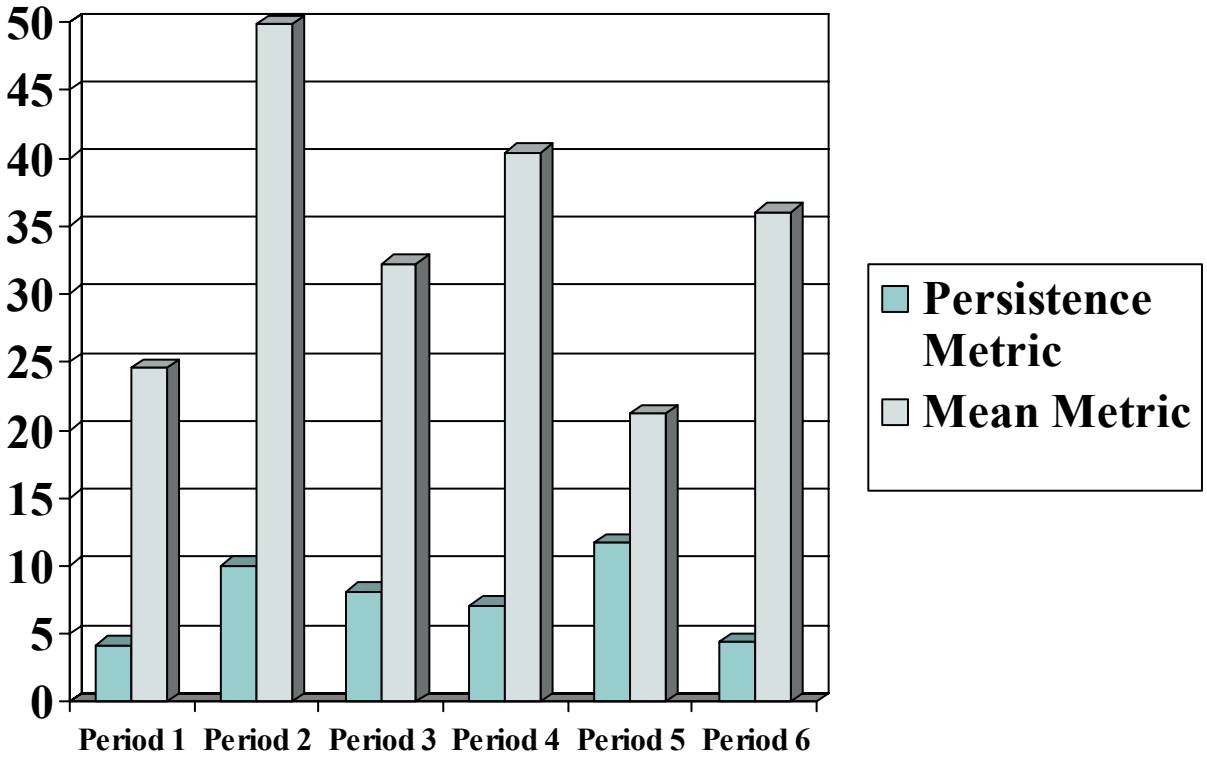
where D_s is for the standard model. In this case, two standard models were used. One standard is a persistence metric which uses the previous measurement as the prediction for the current time step. The second standard is the mean for the entire Carrington rotation.

  - The score is then scaled so that the score is between 0 and 100 by the following transformation $S_i = 50 * (2^{Mi})$

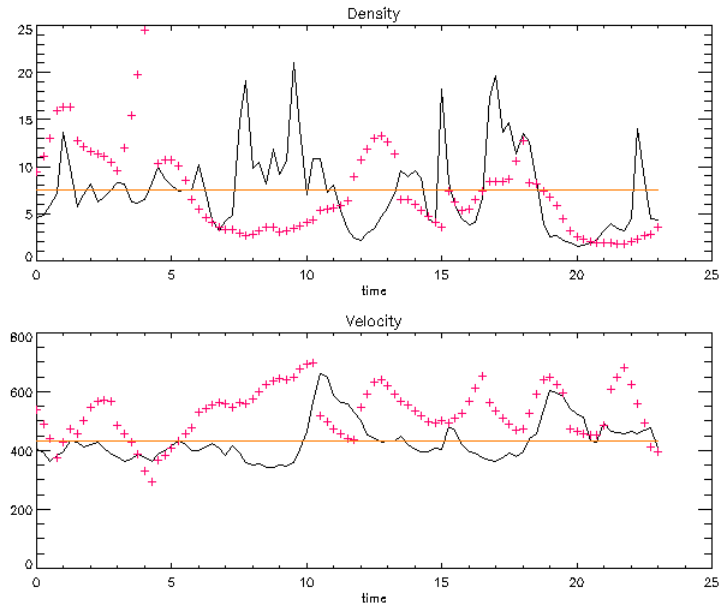
Scores for Density



Scores for Velocity

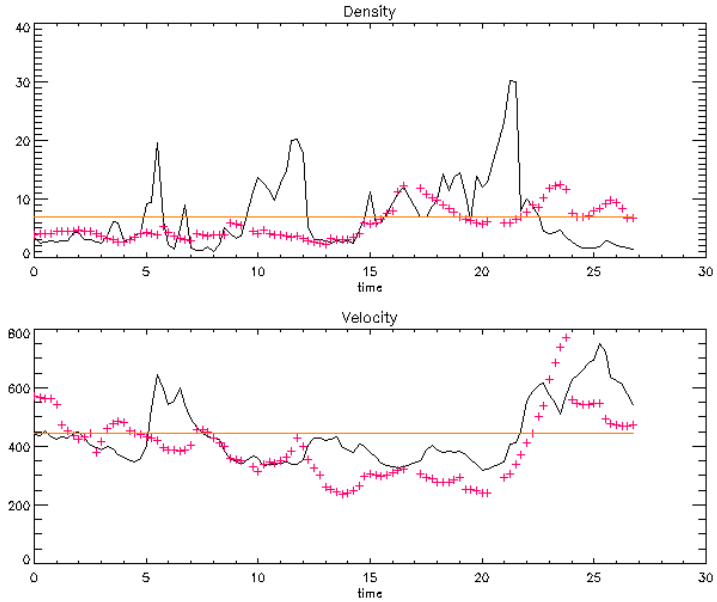


Model and Data Comparison



Time (Days)

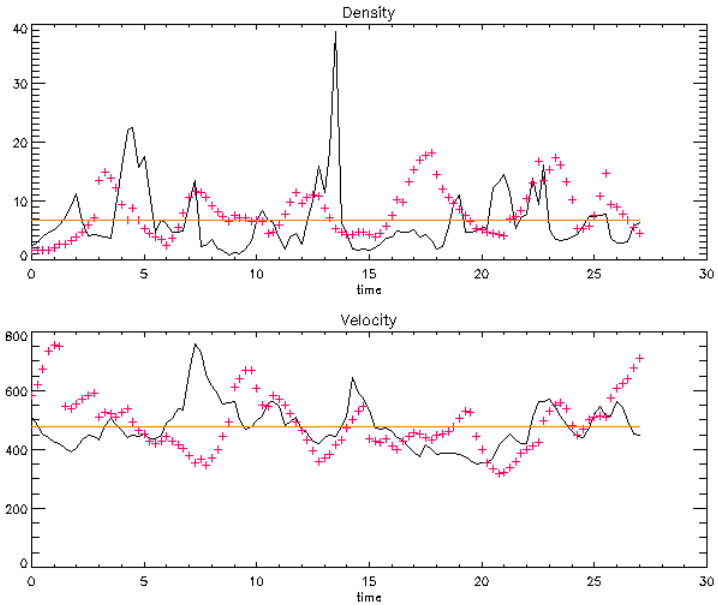
Period 1



Time (Days)

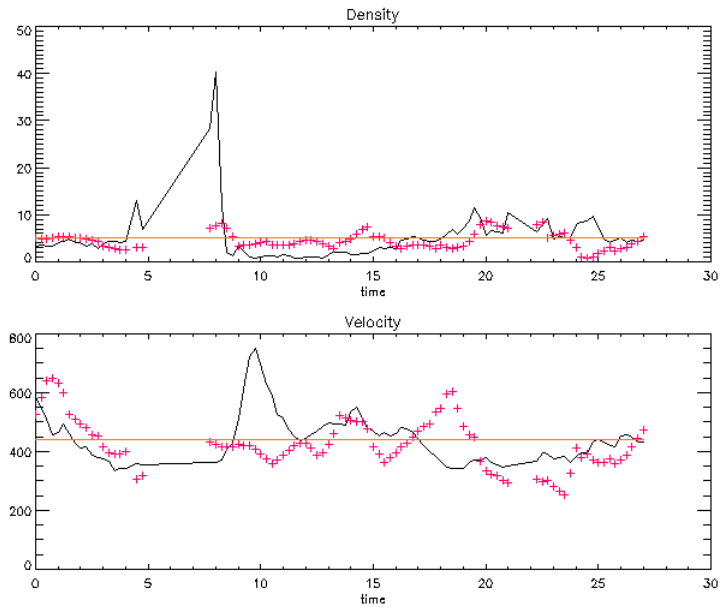
Period 2

Model and Data Comparison



Time (Days)

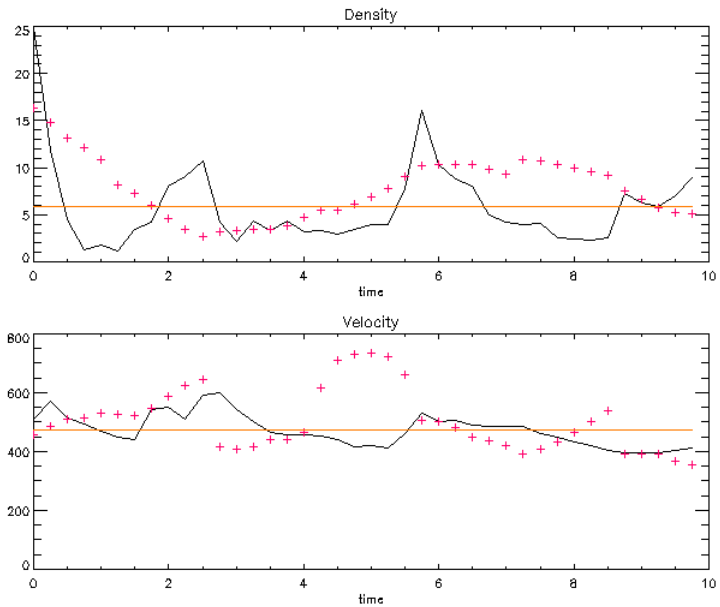
Period 3



Time (Days)

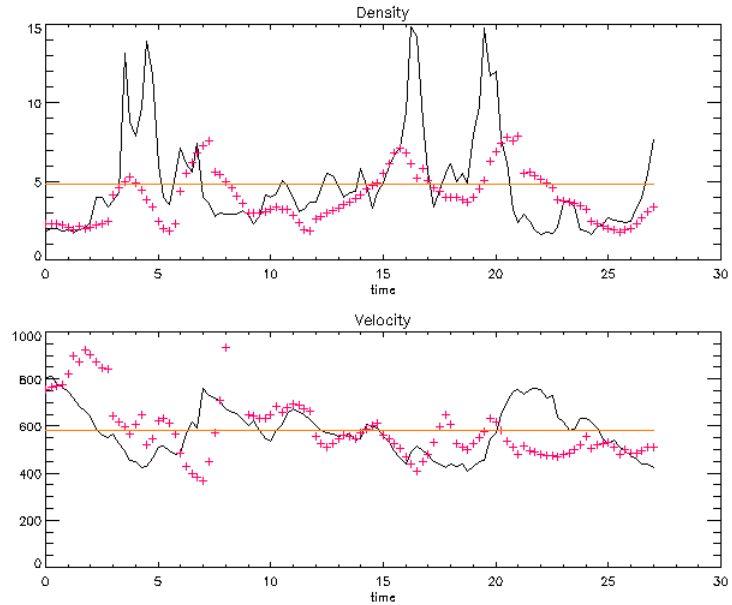
Period 4

Model and Data Comparison



Time (Days)

Period 5



Time (Days)

Period 6

Only 38% of the period covered

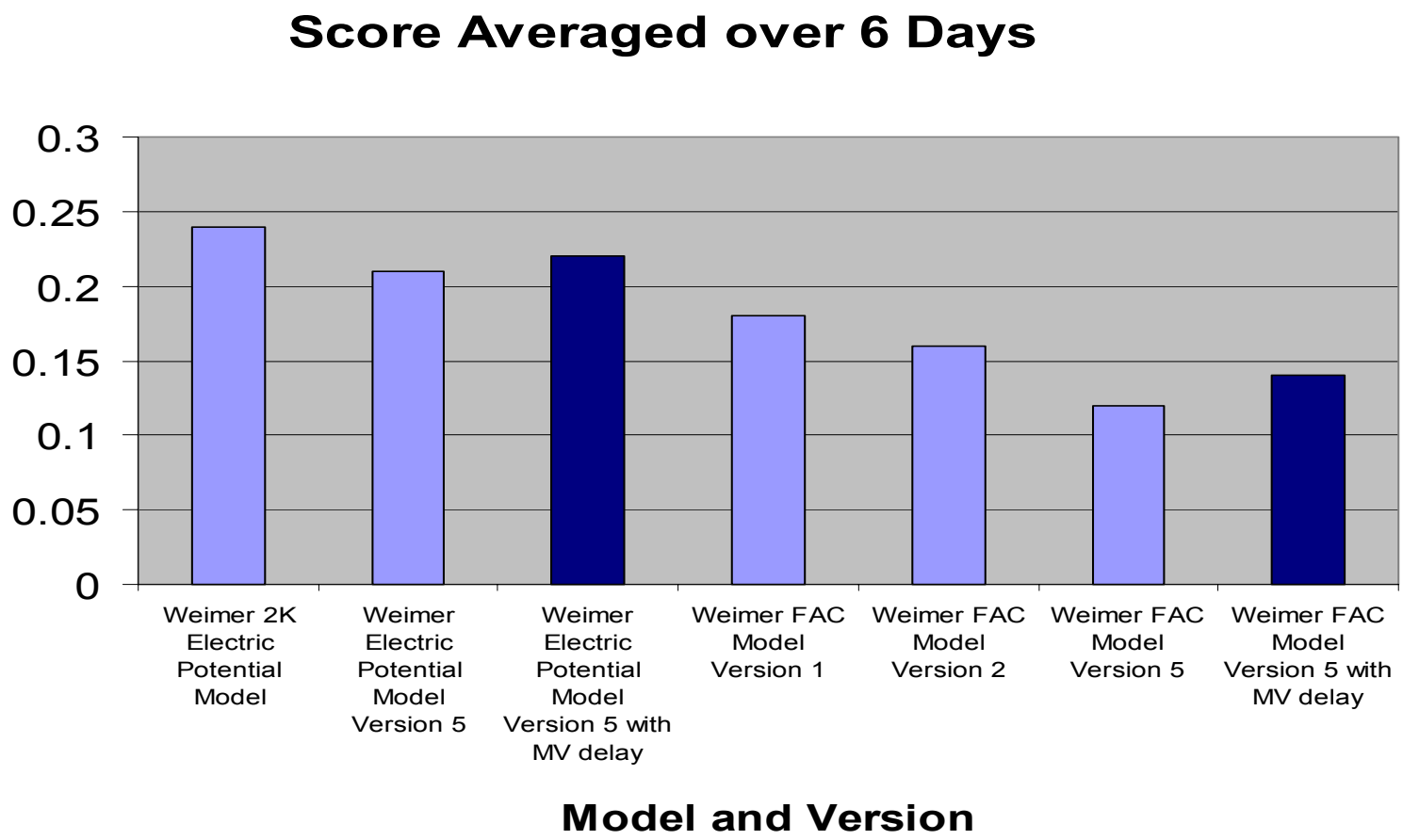
Ground Magnetic Perturbations

- Data
  - 10 stations in the Greenland chain using the H component of the data.
- Models
  - Weimer electric potential model (2 different versions).
  - Weimer field-aligned current model (3 different versions).
- Skill score
  - An individual model is scored $D_i = \Sigma |\Delta H_{model} - \Delta H_{data}| / npts$.
  - A skill score is computed for each ground station by

$$M_i = 1 - D_i / D_s$$

where D_s is for the standard model. In this case, the standard model is $\Delta H_{standard} \equiv 0$.

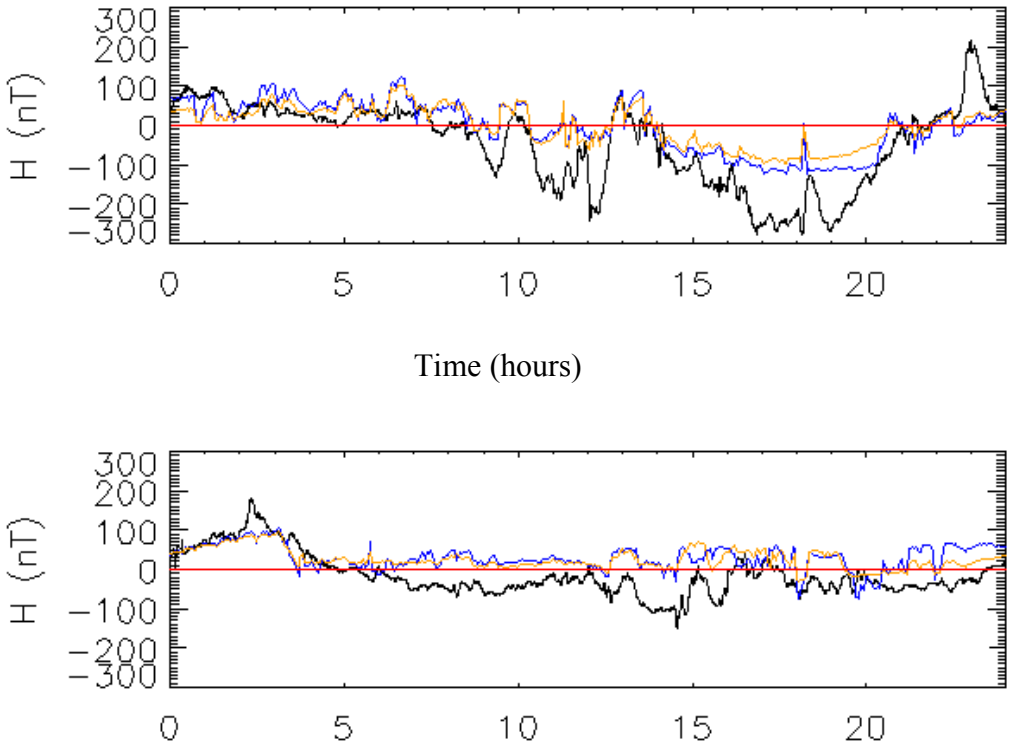
Results for Weimer Models (averaged over 10 stations) for H component.



Parameter Tests

- Different time delays for the ACE data were used. The skill scores were not very sensitive to the time delays. There was a slight improvement when using minimum variance technique received from Dan Weimer.
- Different Hall conductivities were used for the electric potential model. The skill scores were better for Hall conductivities of 5 and 7.5 mhos. For later versions, the scores are more sensitive to different conductivities.

Comparison of Model Results to Data



Black: Data from ground magnetometers

Orange: Model results from Weimer 2k Electric Potential Model

Blue: Model results from Weimer Electric Potential Model Version 5

Magnetometer data was provided by the Danish Meteorological Institute (Dr. Jurgen Watermann, Project Scientist)

Comparison of Model Results to Data Discussion

- In the top plot, the results from the Weimer 2K electric potential model tend to be smaller in magnitude than the results from Weimer electric potential model version 5. Since the results have the same sign as the data, the score for the version 5 model is better for this station on this day. Both scores are in the .2 -.3 range.
- In the bottom plot, the results from the 2K version again tend to be smaller in magnitude than the results from version 5 model. On this day, there is significant periods of time when the model has the wrong sign compared to the data. In this case, the score for the 2K version is better. The scores for this station and day are either negative or around zero.
- For each day, there is at least one station with the wrong sign for a significant period of time. Since the 2K version tends to predict smaller magnitudes, it tends to do better when the sign is incorrect. This tends to give better scores for the 2K version when the scores are averaged over 10 stations.

Proton Fluxes

- Data
  - Proton fluxes from LANL geosynchronous satellites
- Model
  - Fok ring current model driven by MHD models

Skill Scores Determination as Defined by SMC

- **Skill Score using the Log Mean Square Error**
  - Calculate log (mean square error)

    $\text{LogRMSE} = \text{sqrt}\{\sum(\log_{10}|\text{predicted} - \text{observed}|)^2/\text{npts}\}$
  - Calculate log (variance of observations)

    $\text{Log STD} = \text{sqrt}\{\sum(\log_{10}|\text{observed} - \text{mean}|)^2/\text{npts}\}$
  - Skill score

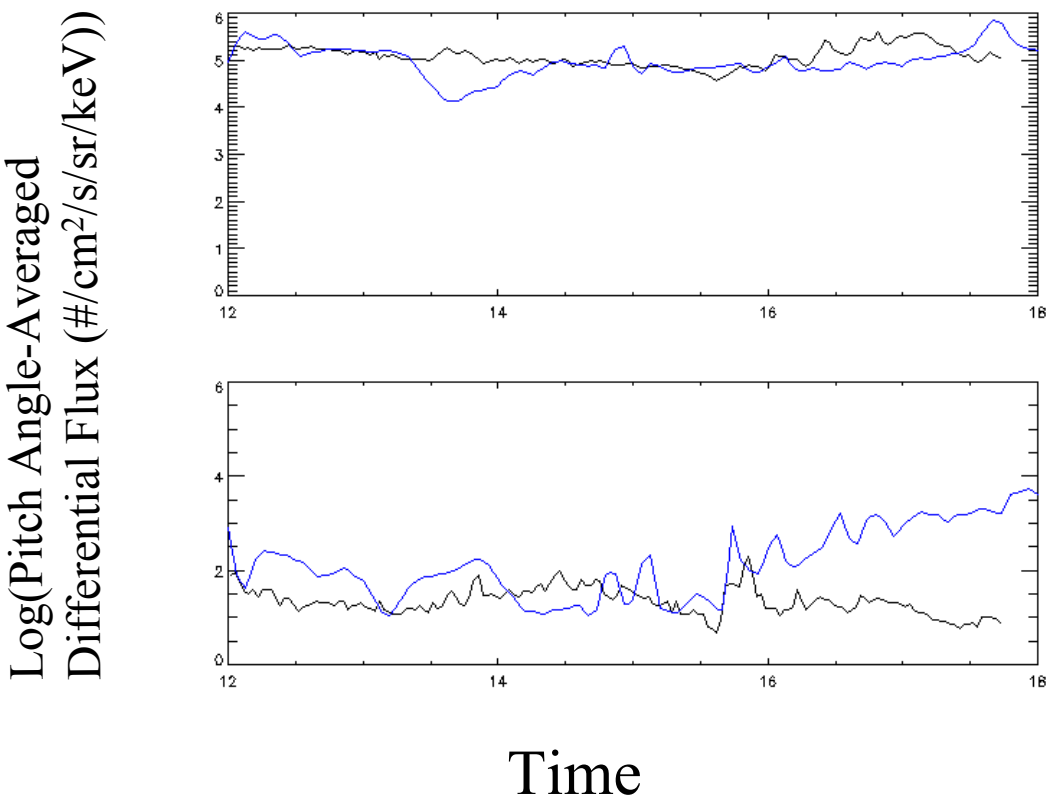
    $\text{Skill score} = 1 - \text{logMSE/logSTD}$
- **Skill Score using the Root Mean Square Deviation**
  - Calculate mean square error

    $\text{RMS\_deviation} = \text{sqrt}\{\sum(\log_{10}[\text{predicted}/\text{observed}])^2/\text{npts}\}$
  - Calculate variance of observations

    $\text{STD\_deviation} = \text{sqrt}\{\sum(\log_{10}[\text{observed}/\text{mean}])^2/\text{npts}\}$
  - Skill score

    $\text{Skill score} = 1 - \text{RMS\_deviation/STD\_deviation}$

Sample of Ring Current Skill Scores Storm Day

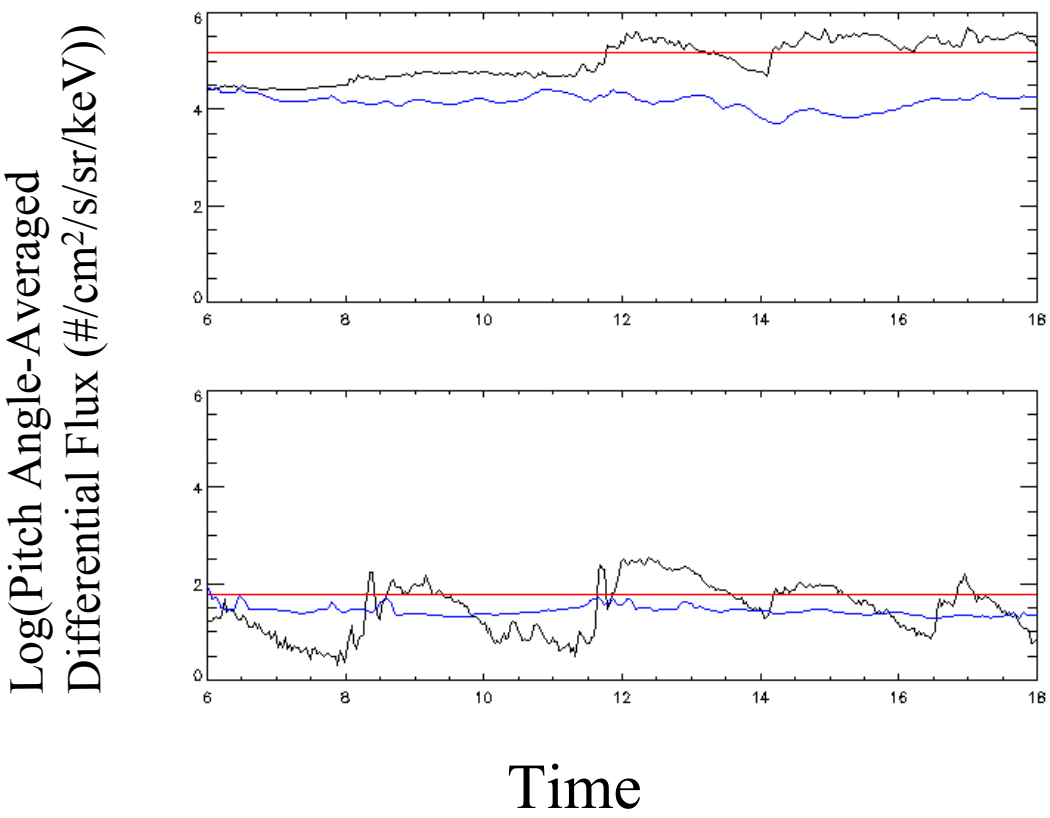


Energy Band (keV)	Log Mean Square	Root Mean Square
50-75	0.0038	.43
250-400	-0.946	-.49

Geosynchronous proton flux data was provided by the Energetic Particle team at Los Alamos National Laboratory, Richard Belian (PI).

Black is LANL data. Blue is the model results.

Sample of Ring Current Skill Scores Sawtooth

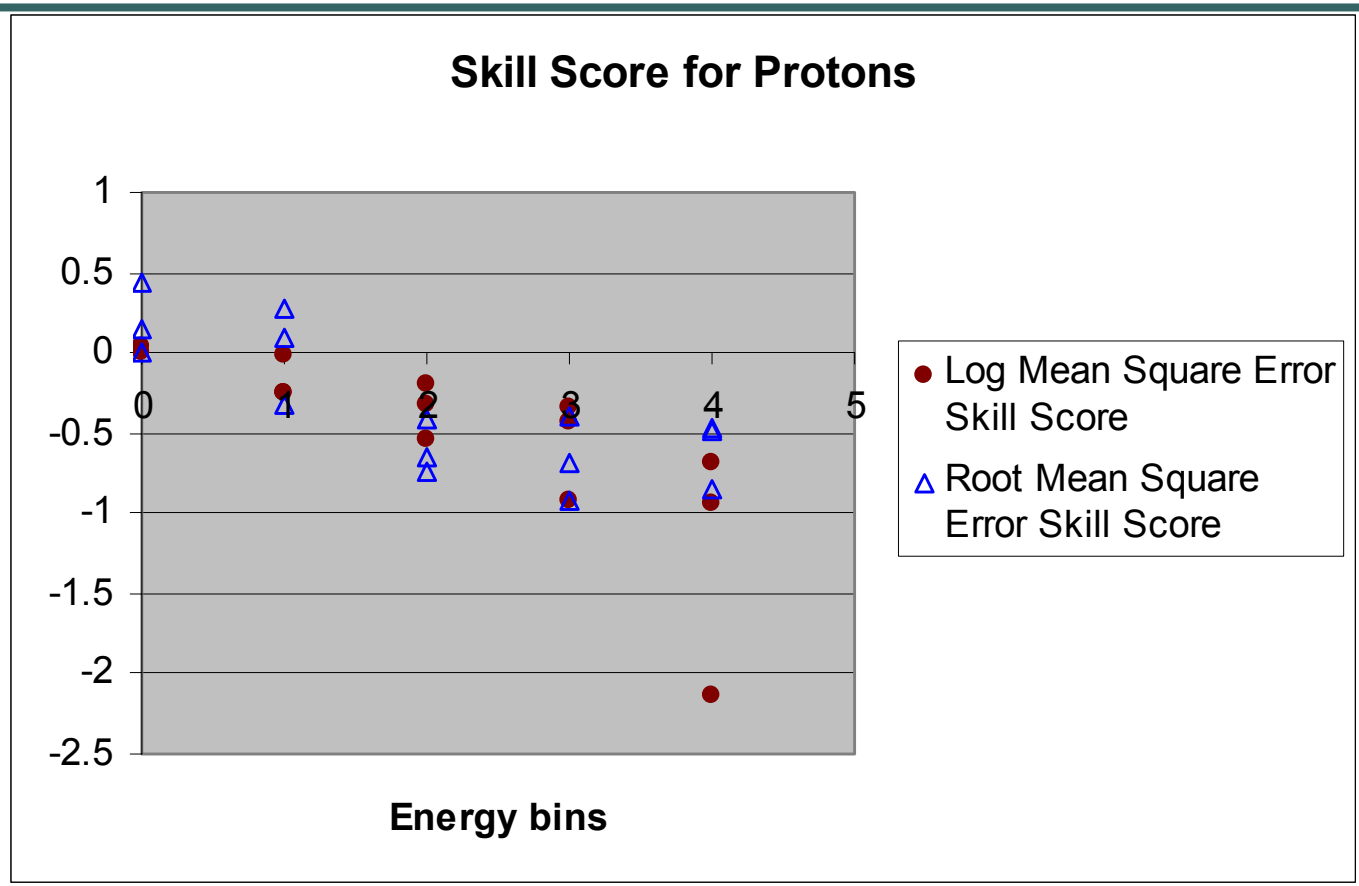


Energy Band (keV)	Log Mean Square	Root Mean Square
50-75	0.0203	-.995
250-400	0.0668	.232

Black is LANL data. Blue is the model results.

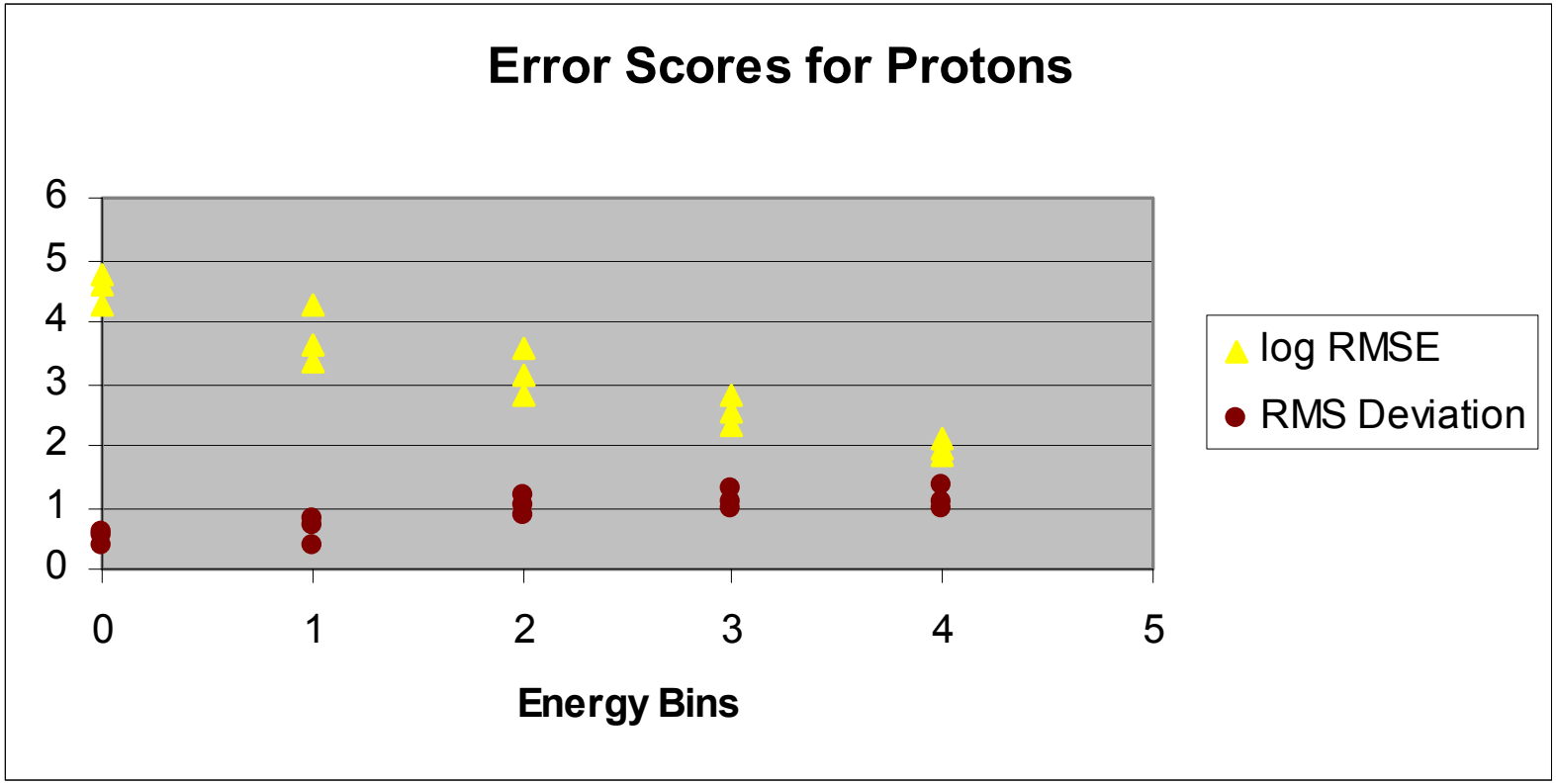
Geosynchronous proton flux data was provided by the Energetic Particle team at Los Alamos National Laboratory, Richard Belian (PI).

Skill Scores by Energy Bin
Storm



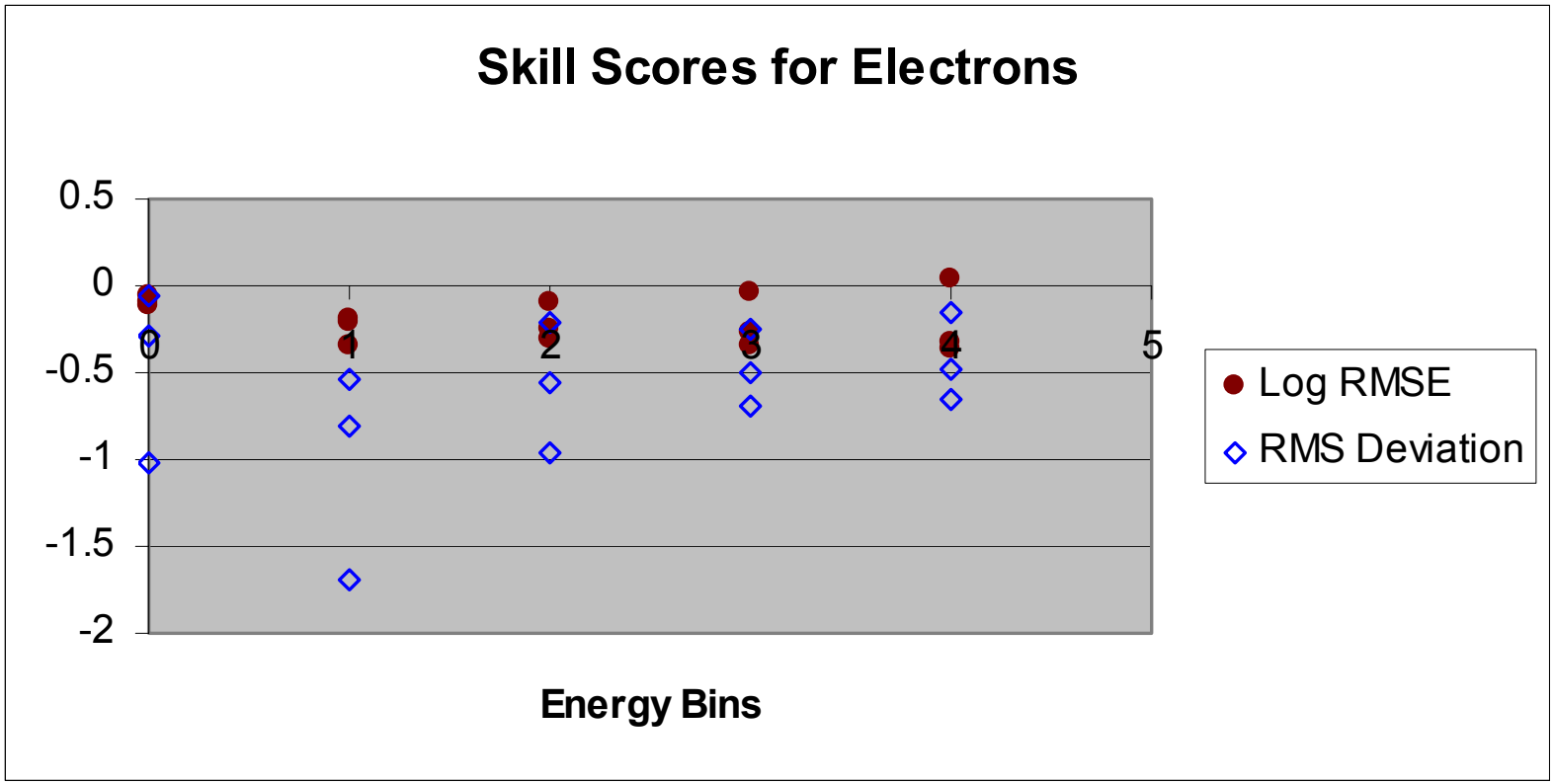
Energy bins (keV): 50-75, 75-113, 113-170, 170-250, 250-400

Error Scores by Energy Bins Storm



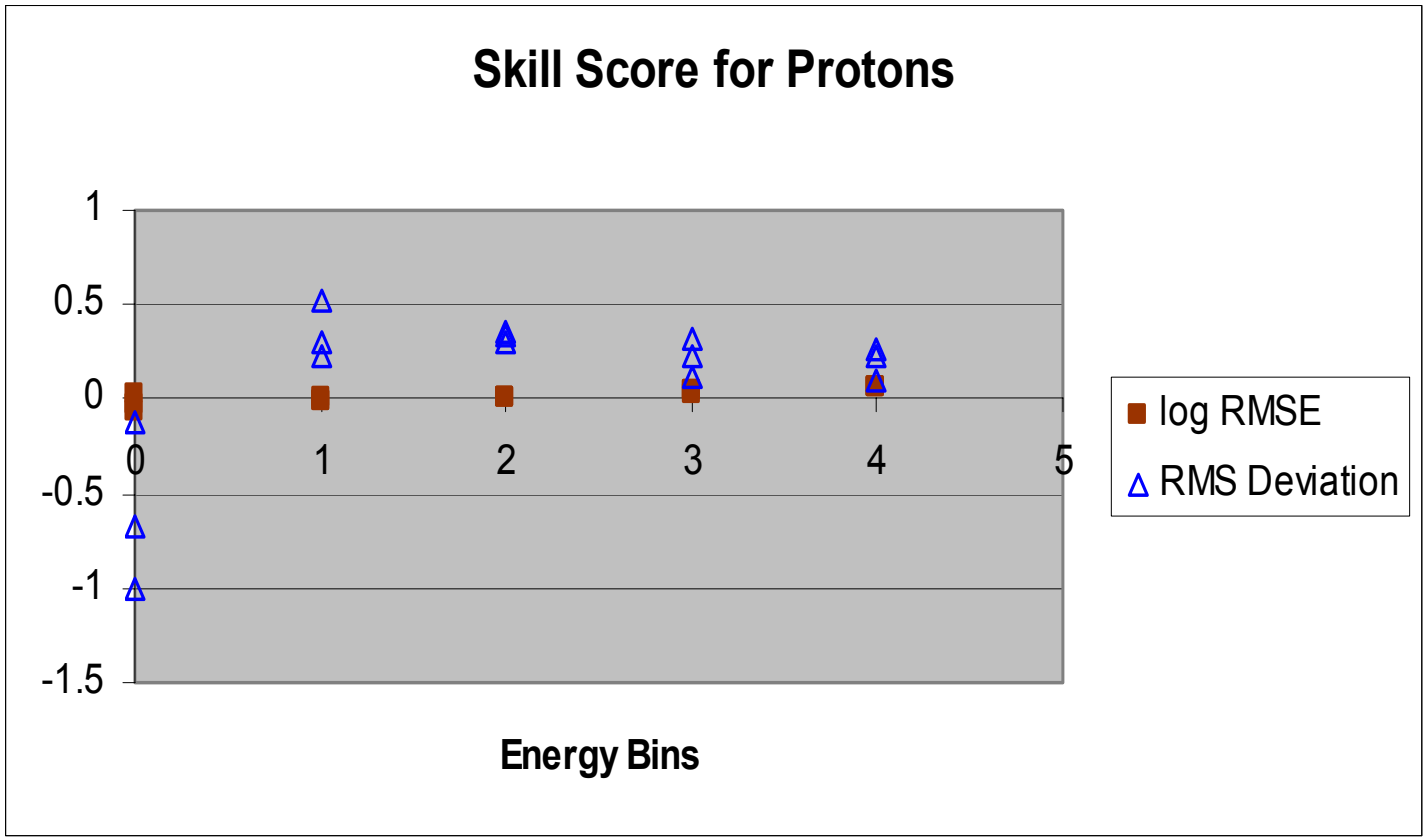
Energy bins (keV): 50-75, 75-113, 113-170, 170-250, 250-400

Skill Scores by Energy Bin
Storm



Energy bins (keV): 50-75, 75-105, 105-150, 150-225, 225-315

Skill Scores by Energy Bin Sawtooth



Energy bins (keV): 50-75, 75-113, 113-170, 170-250, 250-400

Discussion

- The RMS deviation skill score tends to show a larger variation in scores than the Log RMSE skill score.
- On the sawtooth injection day, the model tended to get an “average” flux right but did not see any variation. The RMS deviation skill scores were high for four energy bins (around .2) while the Log RMSE tended to be around 0.
- The skill scores for the electrons were significantly lower.

Other PossibleMetrics for Proton Fluxes

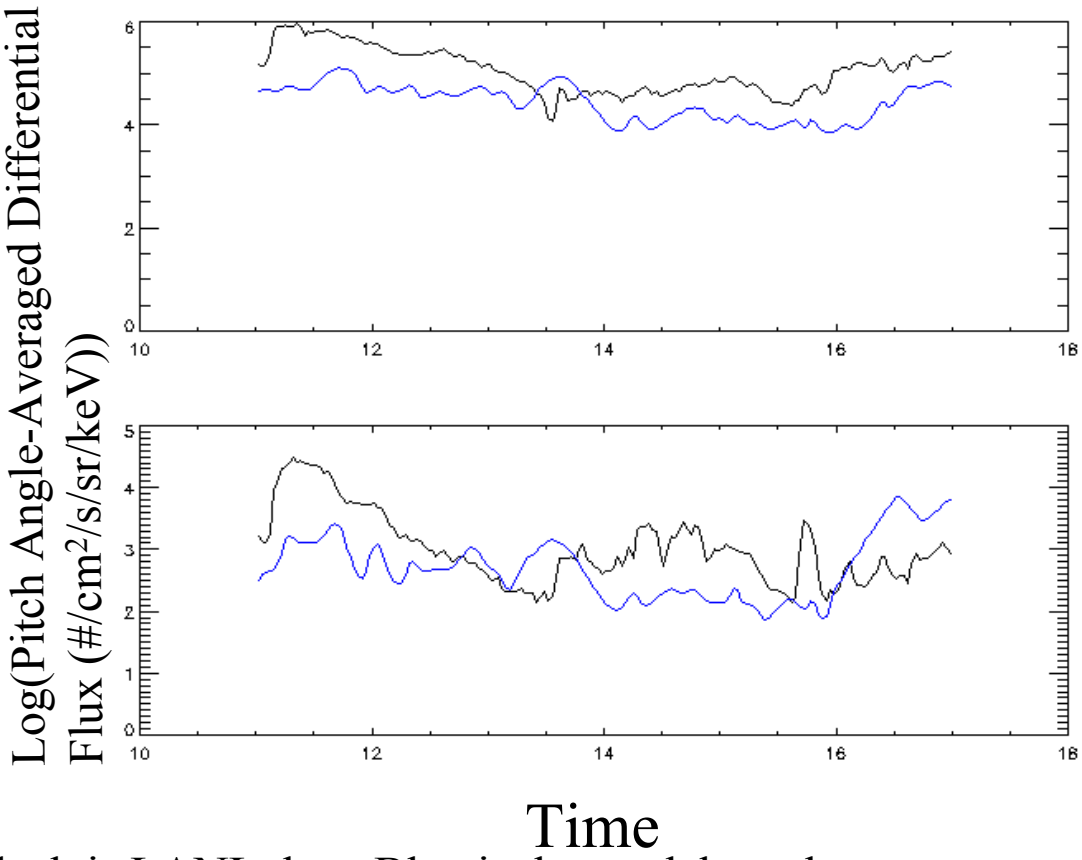
- **Root Mean Square Error Skill Score**
  - Calculate root mean square error (RMSE)

    $RMSE = sqrt(\sum(predicted - observed)^2/npts)$
  - Calculate standard deviation of observations

    $STD = sqrt(\sum(observed - mean)^2/npts)$
  - RMSE skill score

    $Skill\ score = 1 - RMSE/STD$
- **Cross Correlation**

Sample of Ring Current Metric



Black is LANL data. Blue is the model results.

RMSE Score	Skill Cross Correlation
0.07	.59
-0.01	0.07

Geosynchronous proton flux data was provided by the Energetic Particle team at Los Alamos National Laboratory, Richard Belian (PI).

Future Plans for Inner Magnetosphere Models

- We plan to do the skill score using several different energy bands for different days and 2-3 satellites per day.
- We will do the same comparison using electron data at the same energies. In this case, we will test two different versions of the Fok ring current model. These models use different density and temperature profiles.
- We will also do comparisons for higher energies with the Fok radiation belt model.

Future Plans for Global MHD Models

- Community wide metrics
  - To be determined by the community
  - Possible candidates
    - Comparison with DMSP satellites
    - Comparison with GOES data

Future Metric Domains

Need community input on metrics for

- Solar models
- Ionosphere models

Summary

- The ground magnetic perturbations is a first attempt at creation and application of a standard and repeatable metric.
- Blind test (no fine tuning)!
- Fine tuning of metrics is required in collaboration with the operational agencies and researchers.
- First steps, more to come.